

*Stochastic Mechanics*  
*Random Media*  
*Signal Processing and Image Synthesis*  
*Mathematical Economics and Finance*  
*Stochastic Optimization*  
*Stochastic Control*

**Applications of  
Mathematics**

*Stochastic Modelling  
and Applied Probability*

**31**

*Edited by*

I. Karatzas  
M. Yor

*Advisory Board*

P. Brémaud  
E. Carlen  
W. Fleming  
D. Geman  
G. Grimmett  
G. Papanicolaou  
J. Scheinkman

**Springer**  
**Science+Business Media, LLC**

# Applications of Mathematics

---

- 1 Fleming/Rishel, **Deterministic and Stochastic Optimal Control** (1975)
- 2 Marchuk, **Methods of Numerical Mathematics**, Second Ed. (1982)
- 3 Balakrishnan, **Applied Functional Analysis**, Second Ed. (1981)
- 4 Borovkov, **Stochastic Processes in Queueing Theory** (1976)
- 5 Liptser/Shiryayev, **Statistics of Random Processes I: General Theory**, Second Ed. (1977)
- 6 Liptser/Shiryayev, **Statistics of Random Processes II: Applications**, Second Ed. (1978)
- 7 Vorob'ev, **Game Theory: Lectures for Economists and Systems Scientists** (1977)
- 8 Shiryayev, **Optimal Stopping Rules** (1978)
- 9 Ibragimov/Rozanov, **Gaussian Random Processes** (1978)
- 10 Wonham, **Linear Multivariable Control: A Geometric Approach**, Third Ed. (1985)
- 11 Hida, **Brownian Motion** (1980)
- 12 Hestenes, **Conjugate Direction Methods in Optimization** (1980)
- 13 Kallianpur, **Stochastic Filtering Theory** (1980)
- 14 Krylov, **Controlled Diffusion Processes** (1980)
- 15 Prabhu, **Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Communication**, Second Ed. (1998)
- 16 Ibragimov/Has'minskii, **Statistical Estimation: Asymptotic Theory** (1981)
- 17 Cesari, **Optimization: Theory and Applications** (1982)
- 18 Elliott, **Stochastic Calculus and Applications** (1982)
- 19 Marchuk/Shaidourov, **Difference Methods and Their Extrapolations** (1983)
- 20 Hijab, **Stabilization of Control Systems** (1986)
- 21 Protter, **Stochastic Integration and Differential Equations** (1990)
- 22 Benveniste/Métivier/Priouret, **Adaptive Algorithms and Stochastic Approximations** (1990)
- 23 Kloeden/Platen, **Numerical Solution of Stochastic Differential Equations** (1992)
- 24 Kushner/Dupuis, **Numerical Methods for Stochastic Control Problems in Continuous Time**, Second Ed. (2001)
- 25 Fleming/Soner, **Controlled Markov Processes and Viscosity Solutions** (1993)
- 26 Baccelli/Brémaud, **Elements of Queueing Theory** (1994)
- 27 Winkler, **Image Analysis, Random Fields, and Dynamic Monte Carlo Methods: An Introduction to Mathematical Aspects** (1994)
- 28 Kalpazidou, **Cycle Representations of Markov Processes** (1995)
- 29 Elliott/Aggoun/Moore, **Hidden Markov Models: Estimation and Control** (1995)
- 30 Hernández-Lerma/Lasserre, **Discrete-Time Markov Control Processes: Basic Optimality Criteria** (1996)
- 31 Devroye/Györfi/Lugosi, **A Probabilistic Theory of Pattern Recognition** (1996)
- 32 Maitra/Sudderth, **Discrete Gambling and Stochastic Games** (1996)

*(continued after index)*

Luc Devroye   László Györfi  
Gábor Lugosi

# A Probabilistic Theory of Pattern Recognition

With 99 Figures



Springer

Luc Devroye  
School of Computer Science  
McGill University  
Montreal, Quebec, H3A 2A7  
Canada

László Györfi  
Gábor Lugosi  
Department of Mathematics and  
Computer Science  
Technical University of Budapest  
Budapest  
Hungary

*Managing Editors*

I. Karatzas  
Department of Statistics  
Columbia University  
New York, NY 10027, USA

M. Yor  
CNRS, Laboratoire de Probabilités  
Université Pierre et Marie Curie  
4, Place Jussieu, Tour 56  
F-75252 Paris Cedex 05, France

---

Mathematics Subject Classification (1991): 68T10, 68T05, 62G07, 62H30

---

Library of Congress Cataloging-in-Publication Data  
Devroye, Luc.

A probabilistic theory of pattern recognition/Luc Devroye,  
László Györfi, Gábor Lugosi.

p. cm.

Includes bibliographical references and index.

I. Pattern perception. 2. Probabilities. I. Györfi, László.  
II. Lugosi, Gábor. III. Title.  
Q327.D5 1996  
003'.52'015192—dc20

95-44633

Printed on acid-free paper.

© 1996 by Springer Science+Business Media New York  
Originally published by Springer-Verlag New York, Inc. in 1996  
Softcover reprint of the hardcover 1st edition 1996

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or here-after developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Francine McNeill; manufacturing supervised by Jeffrey Taub.  
Photocomposed copy prepared using Springer's svsing.sty macro.

9 8 7 6 5 4 3

SPIN 10830936

ISBN 978-1-4612-6877-2

ISBN 978-1-4612-0711-5 (eBook)

DOI 10.1007/978-1-4612-0711-5

# Preface

Life is just a long random walk. Things are created because the circumstances happen to be right. More often than not, creations, such as this book, are accidental. Nonparametric estimation came to life in the fifties and sixties and started developing at a frenzied pace in the late sixties, engulfing pattern recognition in its growth. In the mid-sixties, two young men, Tom Cover and Peter Hart, showed the world that the nearest neighbor rule in all its simplicity was guaranteed to err at most twice as often as the best possible discrimination method. Tom's results had a profound influence on Terry Wagner, who became a Professor at the University of Texas at Austin and brought probabilistic rigor to the young field of nonparametric estimation. Around 1971, Vapnik and Chervonenkis started publishing a revolutionary series of papers with deep implications in pattern recognition, but their work was not well known at the time. However, Tom and Terry had noticed the potential of the work, and Terry asked Luc Devroye to read that work in preparation for his Ph.D. dissertation at the University of Texas. The year was 1974. Luc ended up in Texas quite by accident thanks to a tip by his friend and fellow Belgian Willy Wouters, who matched him up with Terry. By the time Luc's dissertation was published in 1976, pattern recognition had taken off in earnest. On the theoretical side, important properties were still being discovered. In 1977, Stone stunned the nonparametric community by showing that there are nonparametric rules that are convergent for all distributions of the data. This is called distribution-free or universal consistency, and it is what makes nonparametric methods so attractive. Yet, very few researchers were concerned with universal consistency—one notable exception was Laci Györfi, who at that time worked in Budapest amid an energetic group of nonparametric specialists that included Sándor Csibi, József Fritz, and Pál Révész.

So, linked by a common vision, Luc and Laci decided to join forces in the early eighties. In 1982, they wrote six chapters of a book on nonparametric regression function estimation, but these were never published. In fact, the notes are still in drawers in their offices today. They felt that the subject had not matured yet. A book on nonparametric density estimation saw the light in 1985. Unfortunately, as true baby-boomers, neither Luc nor Laci had the time after 1985 to write a text on nonparametric pattern recognition. Enter Gábor Lugosi, who obtained his doctoral degree under Laci's supervision in 1991. Gábor had prepared a set of rough course notes on the subject around 1992 and proposed to coordinate the project—this book—in 1993. With renewed energy, we set out to write the book that we should have written at least ten years ago. Discussions and work sessions were held in Budapest, Montreal, Leuven, and Louvain-La-Neuve. In Leuven, our gracious hosts were Ed van der Meulen and Jan Beirlant, and in Louvain-La-Neuve, we were gastronomically and spiritually supported by Léopold Simar and Irène Gijbels. We thank all of them. New results accumulated, and we had to resist the temptation to publish these in journals. Finally, in May 1995, the manuscript had bloated to such extent that it had to be sent to the publisher, for otherwise it would have become an encyclopedia. Some important unanswered questions were quickly turned into masochistic exercises or wild conjectures. We will explain subject selection, classroom use, chapter dependence, and personal viewpoints in the Introduction. We do apologize, of course, for all remaining errors.

We were touched, influenced, guided, and taught by many people. Terry Wagner's rigor and taste for beautiful nonparametric problems have infected us for life. We thank our past and present coauthors on nonparametric papers, Alain Berlines, Michel Broniatowski, Ricardo Cao, Paul Deheuvels, András Faragó, Adam Krzyżak, Tamás Linder, Andrew Nobel, Mirek Pawlak, Igor Vajda, Harro Walk, and Ken Zeger. Tamás Linder read most of the book and provided invaluable feedback. His help is especially appreciated. Several chapters were critically read by students in Budapest. We thank all of them, especially András Antos, Miklós Csűrös, Balázs Kégl, István Páli, and Márta Pintér. Finally, here is an alphabetically ordered list of friends who directly or indirectly contributed to our knowledge and love of nonparametrics: Andrew and Roger Barron, Denis Bosq, Prabhir Burman, Tom Cover, Antonio Cuevas, Pierre Devijver, Ricardo Fraiman, Ned Glick, Wenceslao Gonzalez-Manteiga, Peter Hall, Eiichi Isogai, Ed Mack, Arthur Nádas, Georg Pflug, George Roussas, Winfried Stute, Tamás Szabados, Godfried Toussaint, Sid Yakowitz, and Yannis Yatracos.

Gábor diligently typed the entire manuscript and coordinated all contributions. He became quite a  $\text{\TeX}$ pert in the process. Several figures were made by `idraw` and `xfig` by Gábor and Luc. Most of the drawings were directly programmed in PostScript by Luc and an undergraduate student at McGill University, Hisham Petry, to whom we are grateful. For Gábor, this book comes at the beginning of his career. Unfortunately, the other two authors are not so lucky. As both Luc and Laci felt that they would probably not write another book on nonparametric pattern recognition—the random walk must go on—they decided to put their general

view of the subject area on paper while trying to separate the important from the irrelevant. Surely, this has contributed to the length of the text.

So far, our random excursions have been happy ones. Coincidentally, Luc is married to Bea, the most understanding woman in the world, and happens to have two great daughters, Natasha and Birgit, who do not stray off their random courses. Similarly, Laci has an equally wonderful wife, Kati, and two children with steady compasses, Kati and János. During the preparations of this book, Gábor met a wonderful girl, Arrate. They have recently decided to tie their lives together.

On the less amorous and glamorous side, we gratefully acknowledge the research support of NSERC CANADA, FCAR QUEBEC, OTKA HUNGARY, and the exchange program between the Hungarian Academy of Sciences and the Royal Belgian Academy of Sciences. Early versions of this text were tried out in some classes at the Technical University of Budapest, Katholieke Universiteit Leuven, Universität Stuttgart, and Université Montpellier II. We would like to thank those students for their help in making this a better book.

Montreal, Quebec, Canada  
Budapest, Hungary  
Budapest, Hungary

Luc Devroye  
Laci Györfi  
Gábor Lugosi

# Contents

<b>Preface</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The Bayes Error</b>	<b>9</b>
2.1 The Bayes Problem	9
2.2 A Simple Example	11
2.3 Another Simple Example	12
2.4 Other Formulas for the Bayes Risk	14
2.5 Plug-In Decisions	15
2.6 Bayes Error Versus Dimension	17
Problems and Exercises	18
<b>3 Inequalities and Alternate Distance Measures</b>	<b>21</b>
3.1 Measuring Discriminatory Information	21
3.2 The Kolmogorov Variational Distance	22
3.3 The Nearest Neighbor Error	22
3.4 The Bhattacharyya Affinity	23
3.5 Entropy	25
3.6 Jeffreys' Divergence	27
3.7 $F$ -Errors	28
3.8 The Mahalanobis Distance	30
3.9 $f$ -Divergences	31
Problems and Exercises	35



<b>4</b>	<b>Linear Discrimination</b>	<b>39</b>
4.1	Univariate Discrimination and Stoller Splits	40
4.2	Linear Discriminants	44
4.3	The Fisher Linear Discriminant	46
4.4	The Normal Distribution	47
4.5	Empirical Risk Minimization	49
4.6	Minimizing Other Criteria	54
	Problems and Exercises	56
<b>5</b>	<b>Nearest Neighbor Rules</b>	<b>61</b>
5.1	Introduction	61
5.2	Notation and Simple Asymptotics	63
5.3	Proof of Stone's Lemma	66
5.4	The Asymptotic Probability of Error	69
5.5	The Asymptotic Error Probability of Weighted Nearest Neighbor Rules	71
5.6	$k$ -Nearest Neighbor Rules: Even $k$	74
5.7	Inequalities for the Probability of Error	75
5.8	Behavior When $L^*$ Is Small	78
5.9	Nearest Neighbor Rules When $L^* = 0$	80
5.10	Admissibility of the Nearest Neighbor Rule	81
5.11	The $(k, l)$ -Nearest Neighbor Rule	81
	Problems and Exercises	83
<b>6</b>	<b>Consistency</b>	<b>91</b>
6.1	Universal Consistency	91
6.2	Classification and Regression Estimation	92
6.3	Partitioning Rules	94
6.4	The Histogram Rule	95
6.5	Stone's Theorem	97
6.6	The $k$ -Nearest Neighbor Rule	100
6.7	Classification Is Easier Than Regression Function Estimation	101
6.8	Smart Rules	106
	Problems and Exercises	107
<b>7</b>	<b>Slow Rates of Convergence</b>	<b>111</b>
7.1	Finite Training Sequence	111
7.2	Slow Rates	113
	Problems and Exercises	118
<b>8</b>	<b>Error Estimation</b>	<b>121</b>
8.1	Error Counting	121
8.2	Hoeffding's Inequality	122
8.3	Error Estimation Without Testing Data	124
8.4	Selecting Classifiers	125

8.5	Estimating the Bayes Error	128
	Problems and Exercises	129
<b>9</b>	<b>The Regular Histogram Rule</b>	<b>133</b>
9.1	The Method of Bounded Differences	133
9.2	Strong Universal Consistency	138
	Problems and Exercises	142
<b>10</b>	<b>Kernel Rules</b>	<b>147</b>
10.1	Consistency	149
10.2	Proof of the Consistency Theorem	153
10.3	Potential Function Rules	159
	Problems and Exercises	161
<b>11</b>	<b>Consistency of the <math>k</math>-Nearest Neighbor Rule</b>	<b>169</b>
11.1	Strong Consistency	170
11.2	Breaking Distance Ties	174
11.3	Recursive Methods	176
11.4	Scale-Invariant Rules	177
11.5	Weighted Nearest Neighbor Rules	178
11.6	Rotation-Invariant Rules	179
11.7	Relabeling Rules	180
	Problems and Exercises	182
<b>12</b>	<b>Vapnik-Chervonenkis Theory</b>	<b>187</b>
12.1	Empirical Error Minimization	187
12.2	Fingering	191
12.3	The Glivenko-Cantelli Theorem	192
12.4	Uniform Deviations of Relative Frequencies from Probabilities	196
12.5	Classifier Selection	199
12.6	Sample Complexity	201
12.7	The Zero-Error Case	202
12.8	Extensions	206
	Problems and Exercises	208
<b>13</b>	<b>Combinatorial Aspects of Vapnik-Chervonenkis Theory</b>	<b>215</b>
13.1	Shatter Coefficients and VC Dimension	215
13.2	Shatter Coefficients of Some Classes	219
13.3	Linear and Generalized Linear Discrimination Rules	224
13.4	Convex Sets and Monotone Layers	226
	Problems and Exercises	229
<b>14</b>	<b>Lower Bounds for Empirical Classifier Selection</b>	<b>233</b>
14.1	Minimax Lower Bounds	234
14.2	The Case $L_C = 0$	234
14.3	Classes with Infinite VC Dimension	238

14.4	The Case $L_C > 0$	239
14.5	Sample Complexity	245
	Problems and Exercises	247
<b>15</b>	<b>The Maximum Likelihood Principle</b>	<b>249</b>
15.1	Maximum Likelihood: The Formats	249
15.2	The Maximum Likelihood Method: Regression Format	250
15.3	Consistency	253
15.4	Examples	256
15.5	Classical Maximum Likelihood: Distribution Format	260
	Problems and Exercises	261
<b>16</b>	<b>Parametric Classification</b>	<b>263</b>
16.1	Example: Exponential Families	266
16.2	Standard Plug-In Rules	267
16.3	Minimum Distance Estimates	270
16.4	Empirical Error Minimization	275
	Problems and Exercises	276
<b>17</b>	<b>Generalized Linear Discrimination</b>	<b>279</b>
17.1	Fourier Series Classification	280
17.2	Generalized Linear Classification	285
	Problems and Exercises	287
<b>18</b>	<b>Complexity Regularization</b>	<b>289</b>
18.1	Structural Risk Minimization	290
18.2	Poor Approximation Properties of VC Classes	297
18.3	Simple Empirical Covering	297
	Problems and Exercises	300
<b>19</b>	<b>Condensed and Edited Nearest Neighbor Rules</b>	<b>303</b>
19.1	Condensed Nearest Neighbor Rules	303
19.2	Edited Nearest Neighbor Rules	309
19.3	Sieves and Prototypes	309
	Problems and Exercises	312
<b>20</b>	<b>Tree Classifiers</b>	<b>315</b>
20.1	Invariance	318
20.2	Trees with the $X$ -Property	319
20.3	Balanced Search Trees	322
20.4	Binary Search Trees	326
20.5	The Chronological $k$ -d Tree	328
20.6	The Deep $k$ -d Tree	332
20.7	Quadtrees	333
20.8	Best Possible Perpendicular Splits	334
20.9	Splitting Criteria Based on Impurity Functions	336

20.10	A Consistent Splitting Criterion	340
20.11	BSP Trees	341
20.12	Primitive Selection	343
20.13	Constructing Consistent Tree Classifiers	346
20.14	A Greedy Classifier	348
	Problems and Exercises	357
<b>21</b>	<b>Data-Dependent Partitioning</b>	<b>363</b>
21.1	Introduction	363
21.2	A Vapnik-Chervonenkis Inequality for Partitions	364
21.3	Consistency	368
21.4	Statistically Equivalent Blocks	372
21.5	Partitioning Rules Based on Clustering	377
21.6	Data-Based Scaling	381
21.7	Classification Trees	383
	Problems and Exercises	383
<b>22</b>	<b>Splitting the Data</b>	<b>387</b>
22.1	The Holdout Estimate	387
22.2	Consistency and Asymptotic Optimality	389
22.3	Nearest Neighbor Rules with Automatic Scaling	391
22.4	Classification Based on Clustering	392
22.5	Statistically Equivalent Blocks	393
22.6	Binary Tree Classifiers	394
	Problems and Exercises	395
<b>23</b>	<b>The Resubstitution Estimate</b>	<b>397</b>
23.1	The Resubstitution Estimate	397
23.2	Histogram Rules	399
23.3	Data-Based Histograms and Rule Selection	403
	Problems and Exercises	405
<b>24</b>	<b>Deleted Estimates of the Error Probability</b>	<b>407</b>
24.1	A General Lower Bound	408
24.2	A General Upper Bound for Deleted Estimates	411
24.3	Nearest Neighbor Rules	413
24.4	Kernel Rules	415
24.5	Histogram Rules	417
	Problems and Exercises	419
<b>25</b>	<b>Automatic Kernel Rules</b>	<b>423</b>
25.1	Consistency	424
25.2	Data Splitting	428
25.3	Kernel Complexity	431
25.4	Multiparameter Kernel Rules	435

25.5	Kernels of Infinite Complexity	436
25.6	On Minimizing the Apparent Error Rate	439
25.7	Minimizing the Deleted Estimate	441
25.8	Sieve Methods	444
25.9	Squared Error Minimization	445
	Problems and Exercises	446
<b>26</b>	<b>Automatic Nearest Neighbor Rules</b>	<b>451</b>
26.1	Consistency	451
26.2	Data Splitting	452
26.3	Data Splitting for Weighted NN Rules	453
26.4	Reference Data and Data Splitting	454
26.5	Variable Metric NN Rules	455
26.6	Selection of $k$ Based on the Deleted Estimate	457
	Problems and Exercises	458
<b>27</b>	<b>Hypercubes and Discrete Spaces</b>	<b>461</b>
27.1	Multinomial Discrimination	461
27.2	Quantization	464
27.3	Independent Components	466
27.4	Boolean Classifiers	468
27.5	Series Methods for the Hypercube	470
27.6	Maximum Likelihood	472
27.7	Kernel Methods	474
	Problems and Exercises	474
<b>28</b>	<b>Epsilon Entropy and Totally Bounded Sets</b>	<b>479</b>
28.1	Definitions	479
28.2	Examples: Totally Bounded Classes	480
28.3	Skeleton Estimates	482
28.4	Rate of Convergence	485
	Problems and Exercises	486
<b>29</b>	<b>Uniform Laws of Large Numbers</b>	<b>489</b>
29.1	Minimizing the Empirical Squared Error	489
29.2	Uniform Deviations of Averages from Expectations	490
29.3	Empirical Squared Error Minimization	493
29.4	Proof of Theorem 29.1	494
29.5	Covering Numbers and Shatter Coefficients	496
29.6	Generalized Linear Classification	501
	Problems and Exercises	505
<b>30</b>	<b>Neural Networks</b>	<b>507</b>
30.1	Multilayer Perceptrons	507
30.2	Arrangements	511

30.3	Approximation by Neural Networks	517
30.4	VC Dimension	521
30.5	$L_1$ Error Minimization	526
30.6	The Adaline and Padaline	531
30.7	Polynomial Networks	532
30.8	Kolmogorov-Lorentz Networks and Additive Models	534
30.9	Projection Pursuit	538
30.10	Radial Basis Function Networks	540
	Problems and Exercises	542
<b>31</b>	<b>Other Error Estimates</b>	<b>549</b>
31.1	Smoothing the Error Count	549
31.2	Posterior Probability Estimates	554
31.3	Rotation Estimate	556
31.4	Bootstrap	556
	Problems and Exercises	559
<b>32</b>	<b>Feature Extraction</b>	<b>561</b>
32.1	Dimensionality Reduction	561
32.2	Transformations with Small Distortion	567
32.3	Admissible and Sufficient Transformations	569
	Problems and Exercises	572
<b>Appendix</b>		<b>575</b>
A.1	Basics of Measure Theory	575
A.2	The Lebesgue Integral	576
A.3	Denseness Results	579
A.4	Probability	581
A.5	Inequalities	582
A.6	Convergence of Random Variables	584
A.7	Conditional Expectation	585
A.8	The Binomial Distribution	586
A.9	The Hypergeometric Distribution	589
A.10	The Multinomial Distribution	589
A.11	The Exponential and Gamma Distributions	590
A.12	The Multivariate Normal Distribution	590
<b>Notation</b>		<b>591</b>
<b>References</b>		<b>593</b>
<b>Author Index</b>		<b>619</b>
<b>Subject Index</b>		<b>627</b>